SciSpark: Highly Interactive and Scalable Model Evaluation and Climate Metrics for Scientific Data and Analysis

Completed Technology Project (2015 - 2017)

Project Introduction

We will construct SciSpark, a scalable system for interactive model evaluation and for the rapid development of climate metrics and analyses. SciSpark directly leverages the Apache Spark technology and its notion of Resilient Distributed Datasets (RDDs). RDDs represent an immutable data set that can be reused across multi-stage operations, partitioned across multiple machines and automatically reconstructed if a partition is lost. The RDD notion directly enables the reuse of array data across multi-stage operations and it ensures data can be replicated, distributed and easily reconstructed in different storage tiers, e.g., memory for fast interactivity, SSDs for near real time availability and I/O oriented spinning disk for later operations. RDDs also allow Spark's performance to degrade gracefully when there is not sufficient memory available to the system. It may seem surprising to consider an in-memory solution for massive datasets, however a recent study found that at Facebook 96% of active jobs could have their entire data inputs in memory at the same time. In addition, it is worth noting that Spark has shown to be 100x faster in memory and 10x faster on disk than Apache Hadoop, the de facto industry platform for Big Data. Hadoop scales well and there are emerging examples of its use in NASA climate projects (e.g., Teng et al. and Schnase et al.) but as is being discovered in these projects, Hadoop is most suited for batch processing and long running operations. SciSpark contributes a Scientific RDD that corresponds to a multi-dimensional array representing a scientific measurement subset by space, or by time. Scientific RDDs can be created in a handful of ways by: (1) directly loading HDF and NetCDF data into Hadoop Distributed File System (HDFS); (2) creating a partition or split function that divides up a multi-dimensional array by space or time; (3) taking the results of a regridding operation or a climate metrics computation; or (4) telling SciSpark to cache an existing Scientific RDD (sRDD), keeping it cached in memory for data reuse between stages. Scientific RDDs will form the basis for a variety of advanced and interactive climate analyses, starting by default in memory, and then being cached and replicated to disk when not directly needed. SciSpark will also use the Shark interactive SQL technology that allows structured query language (SQL) to be used to store/retrieve RDDs; and will use Apache Mesos to be a good tenant in cloud environments interoperating with other data system frameworks (e.g., HDFS, iRODS, SciDB, etc.). One of the key components of SciSpark is interactive sRDD visualizations and to accomplish this SciSpark delivers a user interface built around the Data Driven Documents (D3) framework. D3 is an immersive, javascript based technology that exploits the underlying Document Object Model (DOM) structure of the web to create histograms, cartographic displays and inspections of climate variables and statistics. SciSpark is evaluated using several topical iterative scientific algorithms inspired by the NASA RCMES project including machine-learning (ML) based clustering of temperature PDFs and other quantities over North America, and graph-based algorithms for searching for Mesocale Convective Complexes in West Africa.

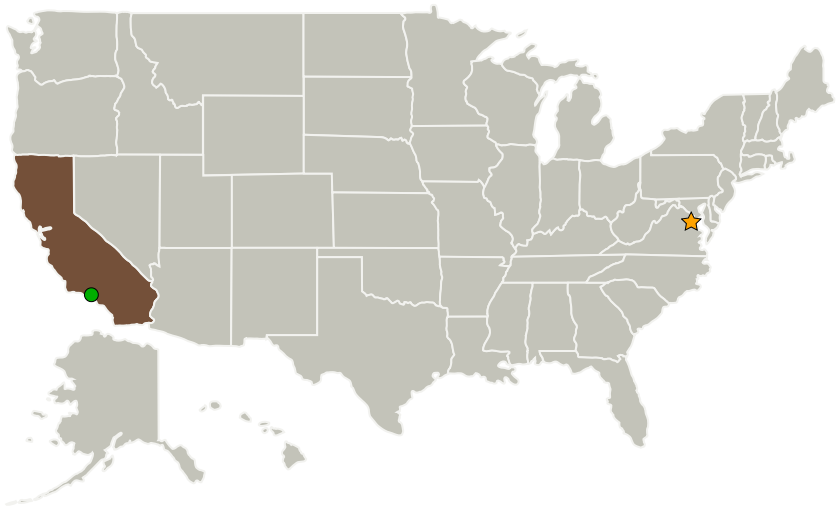
SciSpark: Highly Interactive and Scalable Model Evaluation and Climate Metrics for Scientific Data and Analysis

Table of Contents

SciSpark: Highly Interactive and Scalable Model Evaluation and Climate Metrics for Scientific Data and Analysis

Completed Technology Project (2015 - 2017)

Primary U.S. Work Locations and Key Partners



Organizations Performing Work	Role	Type	Location
★NASA Headquarters(HQ)	Lead Organization	NASA Center	Washington, District of Columbia
●Jet Propulsion Laboratory(JPL)	Supporting Organization	NASA Center	Pasadena, California

Primary U.S. Work Locations
California

Organizational Responsibility

Responsible Mission Directorate:

Science Mission Directorate (SMD)

Lead Center / Facility:

NASA Headquarters (HQ)

Responsible Program:

Advanced Information Systems Technology

Project Management

Program Director:

Pamela S Millar

Program Manager:

Jacqueline J Le Moigne

Principal Investigator:

Christian A Mattmann

Co-Investigators:

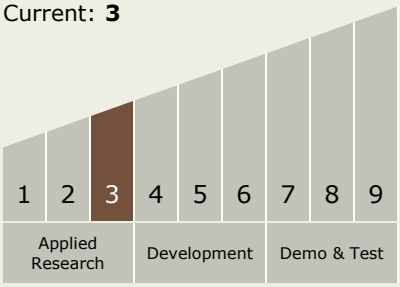
Yolanda Gil
Kim D Whitehall
Huikyo Lee
Jinwon Kim
Paul C Loikith
Duane E Waliser
Karen R Piggee
Brian D Wilson
Lewis J Mcgibbney
Eric J Fetzer

SciSpark: Highly Interactive and Scalable Model Evaluation and Climate Metrics for Scientific Data and Analysis

Completed Technology Project (2015 - 2017)

Technology Maturity (TRL)

Start: **3**
Current: **3**



Technology Areas

Primary:

- TX11 Software, Modeling, Simulation, and Information Processing
  - TX11.4 Information Processing
    - TX11.4.2 Intelligent Data Understanding

Target Destination

Earth